Language Technologies Institute

Carnegie Mellon University

Multimodal Affective Computing

Lecture 2: Measuring Psychological Constructs

Jeffrey Girard
Louis-Philippe Morency

Outline of this week's lecture

- What is a psychological construct?
  - Constructs, indicators, and hierarchies
  - Measurement and construct estimation
- How are constructs commonly measured?
  - Self-report questionnaires
  - Observational and judgment studies
- When is measurement trustworthy?
  - Validity and reliability of measurement
  - An introduction to measurement validation

What is a psychological construct?

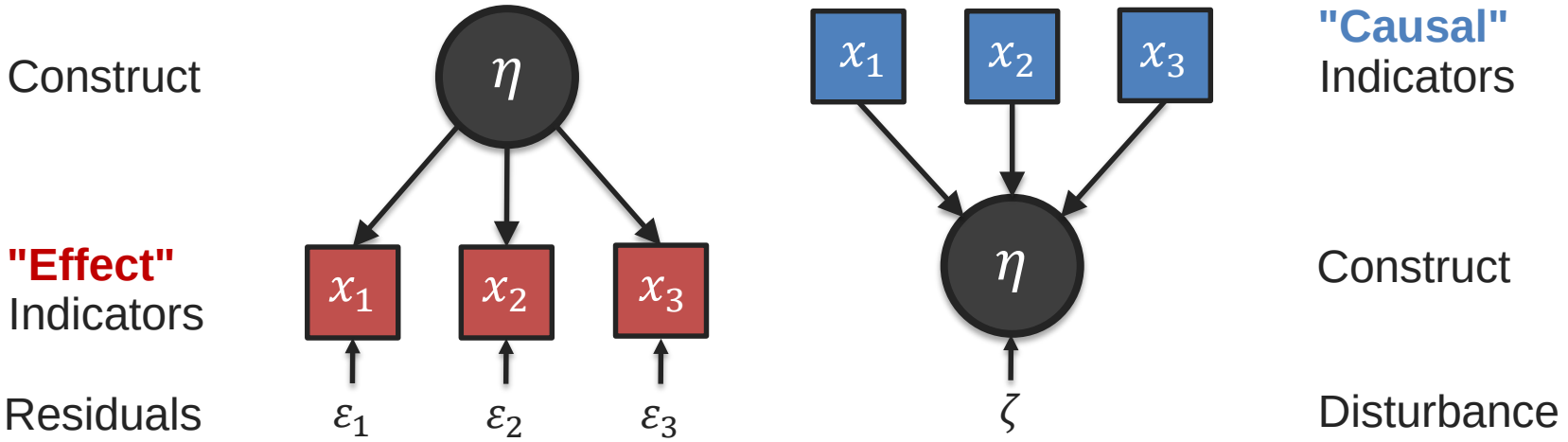
Constructs, indicators, and hierarchies

Constructs, indicators, and hierarchies

- To understand and predict events in the world, it often helps to hypothesize **explanatory latent variables**
- These explanatory variables aren't directly observable but rather **are inferred from observations they explain**
- Examples of explanatory latent variables:
  - *Genes* explain characteristics passing on to offspring
  - *Cancer* explains abnormal cell growth and expansion
  - *Motivation* explains people acting in pursuit of goals

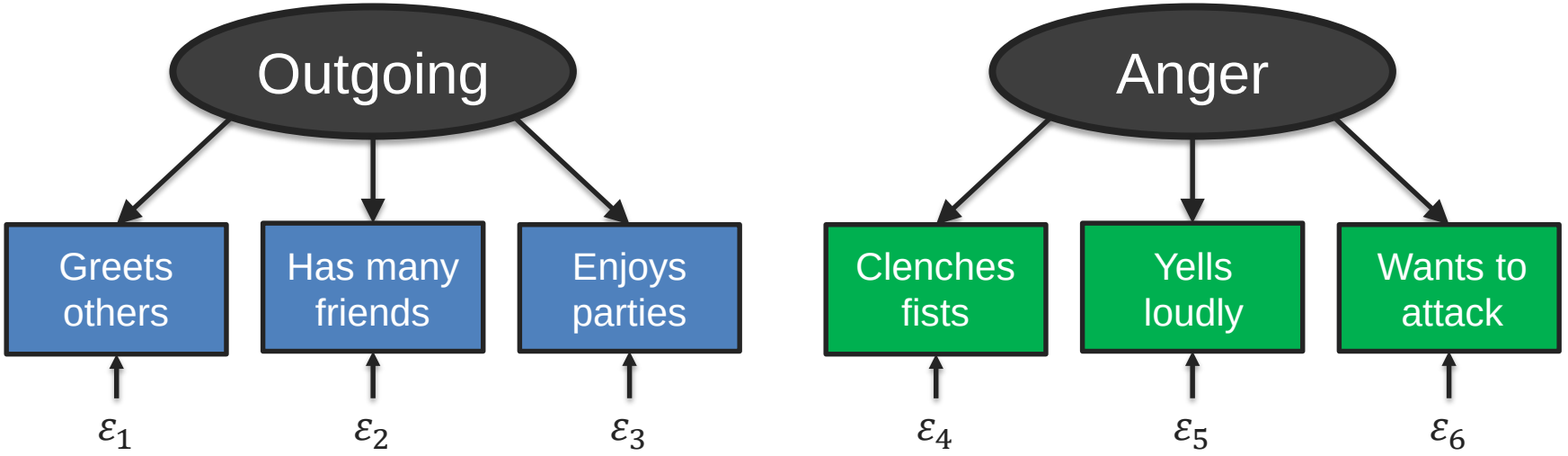
Constructs, indicators, and hierarchies

- Hypothesized latent variables are called **constructs**
- The observations that they explain are called **indicators**
- Most constructs are theorized to **cause** their indicators



Constructs, indicators, and hierarchies

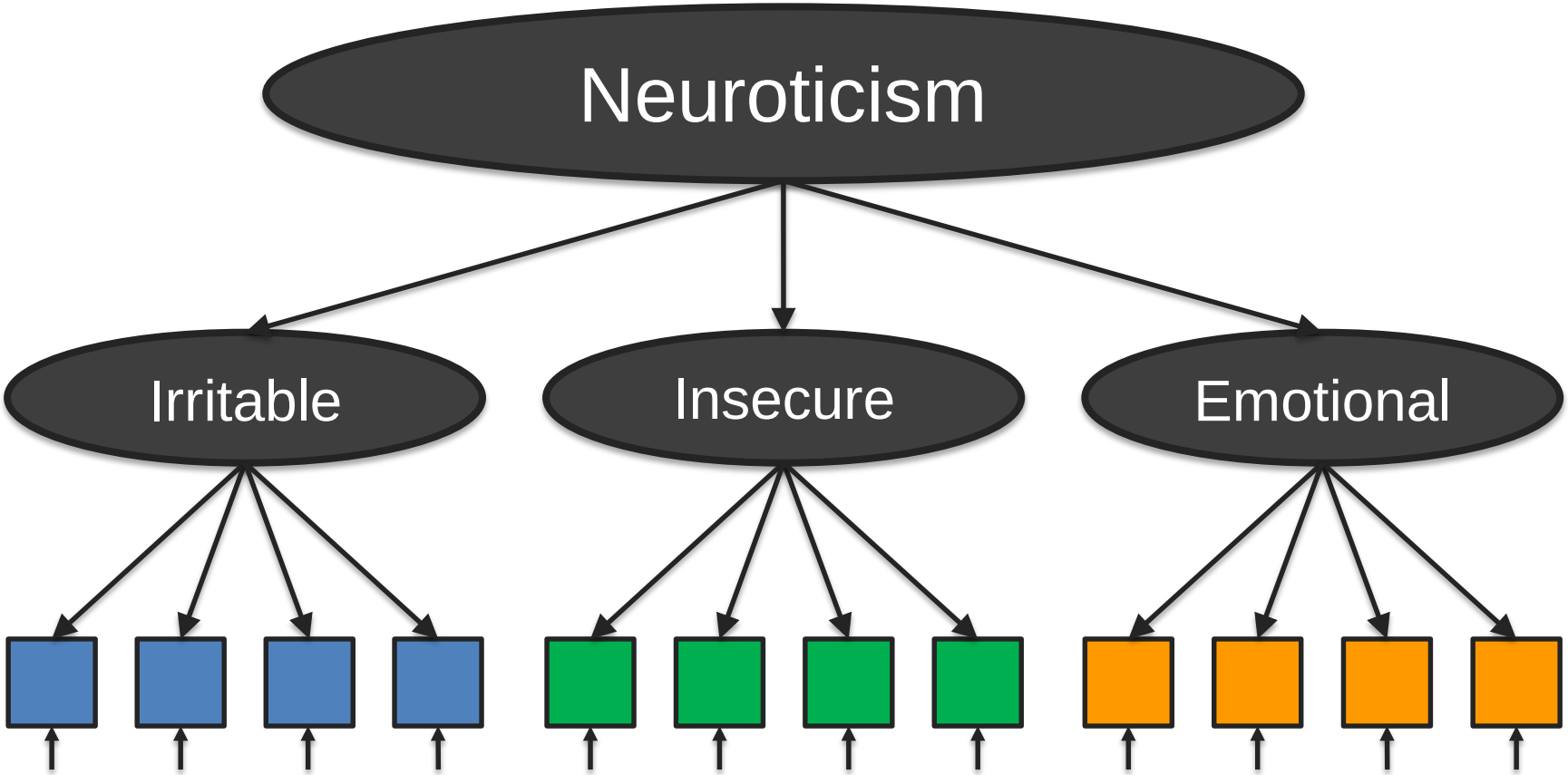
- *Psychological* constructs explain **behavior** and **mind** (and many related internal and external phenomena)
- Indicators can be **internally** experienced phenomena (thoughts/feelings) or **externally** observable (behaviors)



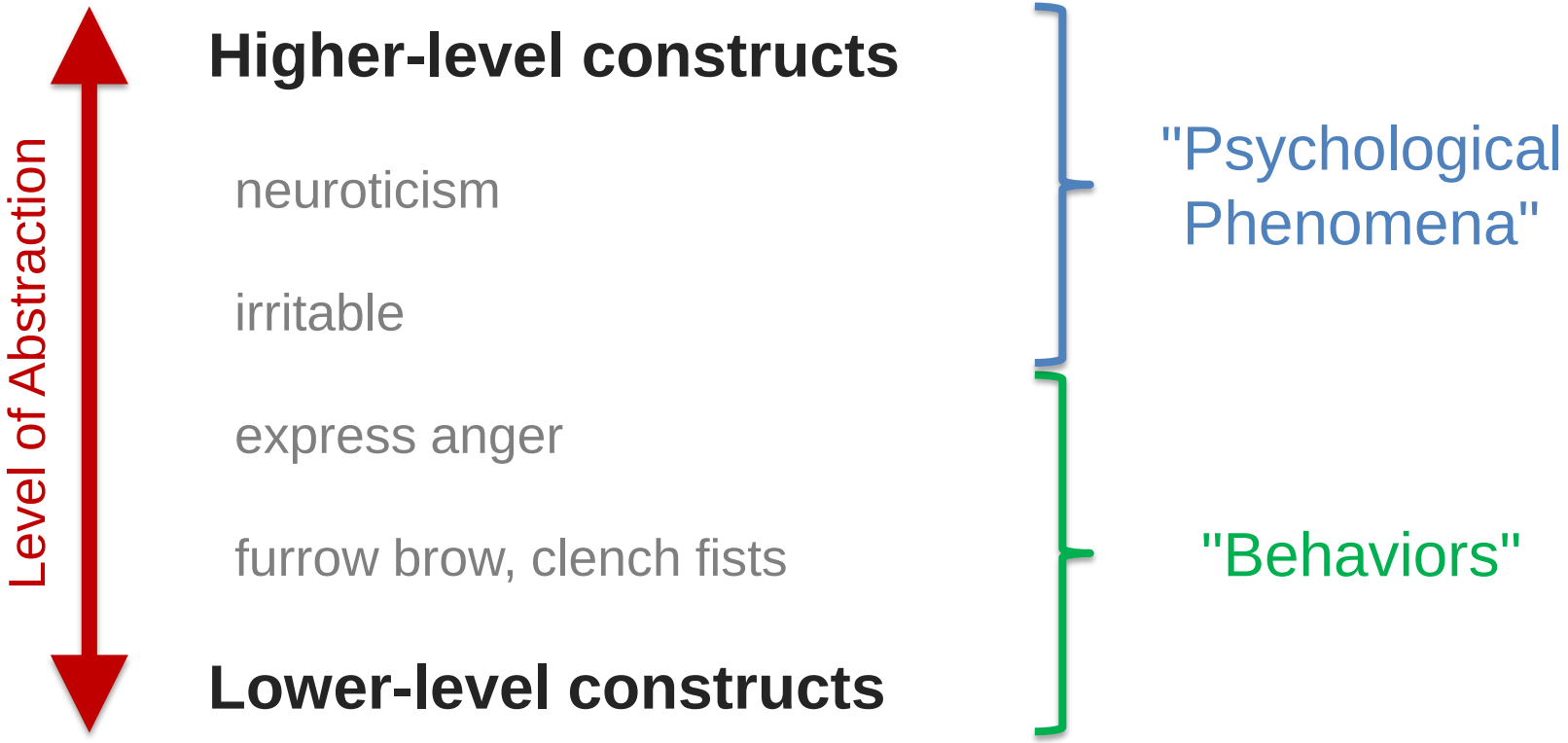
Constructs, indicators, and hierarchies

- Constructs are often **correlated** with other constructs
  - These constructs share similar mechanisms and indicators
  - Some constructs are broader/narrower versions of others
- These relationships often form a construct **hierarchy**
  - Higher-level constructs are more broad, general, and abstract
  - Lower-level constructs are more narrow, specific, and concrete
- Construct hierarchies are very common and **useful**
  - Hierarchies have advanced work in many areas of psychology
  - Hierarchies can improve annotation, modeling, and prediction

Constructs, indicators, and hierarchies



Constructs, indicators, and hierarchies



Measurement and construct estimation

Measurement and construct estimation

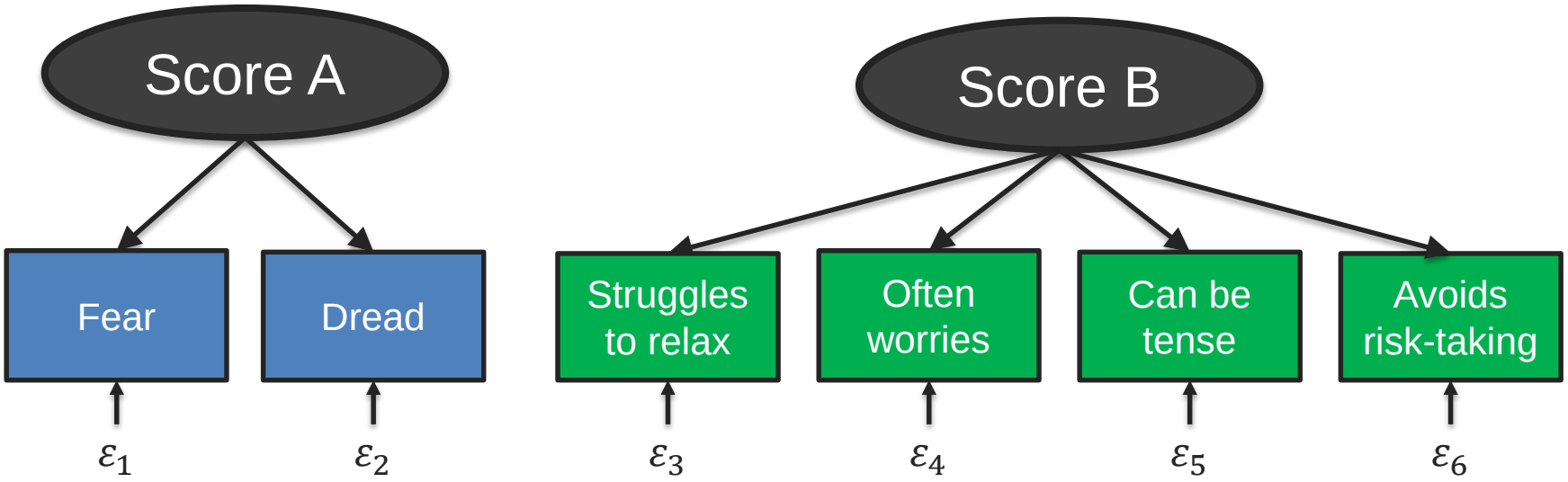
- We often want to know an individual, group, or object's "standing" or **score** on a psychological construct
  - How *neurotic* is a person in their day-to-day life?
  - How *engaged* was the audience during the movie?
  - How *persuasive* is this argument (to most people)?
  - Is the person in the photograph *smiling* or not?
- This means assigning a numerical score to them
- This is called construct **estimation** or **measurement**

Measurement and construct estimation

- But, by definition, constructs cannot be directly measured
- So we must *infer* their scores from measured indicators
- Estimating construct scores thus proceeds as follows:
  1. **Select** a set of indicators to represent the construct
  2. **Measure** each selected indicator for all objects of interest
  3. **Aggregate** these measures into estimated construct scores

Measurement and construct estimation

- What are the pros and cons of scores A and B?
- When might you prefer one score over the other?
- What would be appropriate labels for these scores?



Measurement and construct estimation

Notes on selecting indicators

- Different indicators can yield very different construct estimates
- Use theory and empirical data to inform your indicator selection

Notes on measuring indicators

- Noise and bias in indicator measures can affect construct estimates
- Including multiple indicators can help overcome measurement error

Notes on measure aggregation

- Sums and means are often used but make some big assumptions
- More powerful for aggregation are latent variable models (e.g., SEM)

How are constructs commonly measured?

Self-report questionnaires

Self-report questionnaires

- Self-report questionnaires ask participants to describe themselves on one or more constructs
- Questionnaires are composed of multiple **items**, each of which is responded to using a **rating scale**
- Each item is meant to measure a single indicator
- Construct scores are often estimated by summing or averaging all items that correspond to that construct

Self-report questionnaires

Example: Generalized Anxiety Disorder Scale (GAD-7)

Over the last 2 weeks, how often have you been bothered by the following problems?	Not at all	Several days	Over half the days	Nearly every day
1. Feeling nervous, anxious, or on edge	0	1	**2**	3
2. Not being able to stop or control worrying	0	1	**2**	3
3. Worrying too much about different things	0	1	2	**3**
4. Trouble relaxing	0	**1**	2	3
5. Being so restless that it's hard to sit still	0	**1**	2	3
6. Becoming easily annoyed or irritable	**0**	1	2	3
7. Feeling afraid something awful might happen	**0**	1	2	3

Sum = **9**/21, Mean = **1.29**/3

Self-report questionnaires

- Assumptions of numerical questionnaires
  - All items are measuring the same construct*
  - All items are equally important/central to the construct*
  - All items were correctly understood by the participants
- Pros and cons of self-report questionnaires
  - Common, efficient, and good for internal experiences
  - Can be subjective and influenced by self-report biases

*There are statistical methods to test and relax these assumptions.

Self-report questionnaires

- Final notes on self-report questionnaires
  - Creating a high quality questionnaire is a lot of work
  - It's usually better to use an existing questionnaire*
  - Not all published questionnaires are high quality*
  - Single-item measures tend to be unreliable
- Any questions about self-report questionnaires?

*I will provide some recommendations next week.

Observers and judges

Observers and judges

- Observers and judges are individuals who view stimuli and provide scores on various constructs
- These measurements are standardized using an instrument (e.g., coding scheme or rating scale)
- Such instruments tell observers what to focus on and help them to make consistent measurements
- The goal is to take out any *unwanted* subjectivity

Observers and judges

- **Example: "Simple Smile Coding Scheme"**

*A smile is a facial movement that pulls the mouth corners upwards and toward the ears (see examples below). You will view several images; please examine each image carefully and determine whether the image would be best described as either **Smiling** or **Not Smiling**.*

Examples of **Smiling** Images

Examples of **Not Smiling** Images

Observers and judges

Smiling or **Not Smiling**?

Smiling or **Not Smiling**?

Observers and judges

- How could we improve this coding system (e.g., changes to the instructions, examples, or available categories)?
  - Clarify if smiles that appear awkward or negative should count
  - Add examples of "boundary cases" or more difficult images
  - Add a category for "no face" or "not applicable" or "can't see"
- What other questions could we ask about each image in order to measure other similar/dissimilar constructs?
  - How positive does the person in the image appear to feel?
  - Is the mouth open (or are the lips parted) in the image?
  - Is the face in the image neutral or emotional in some way?

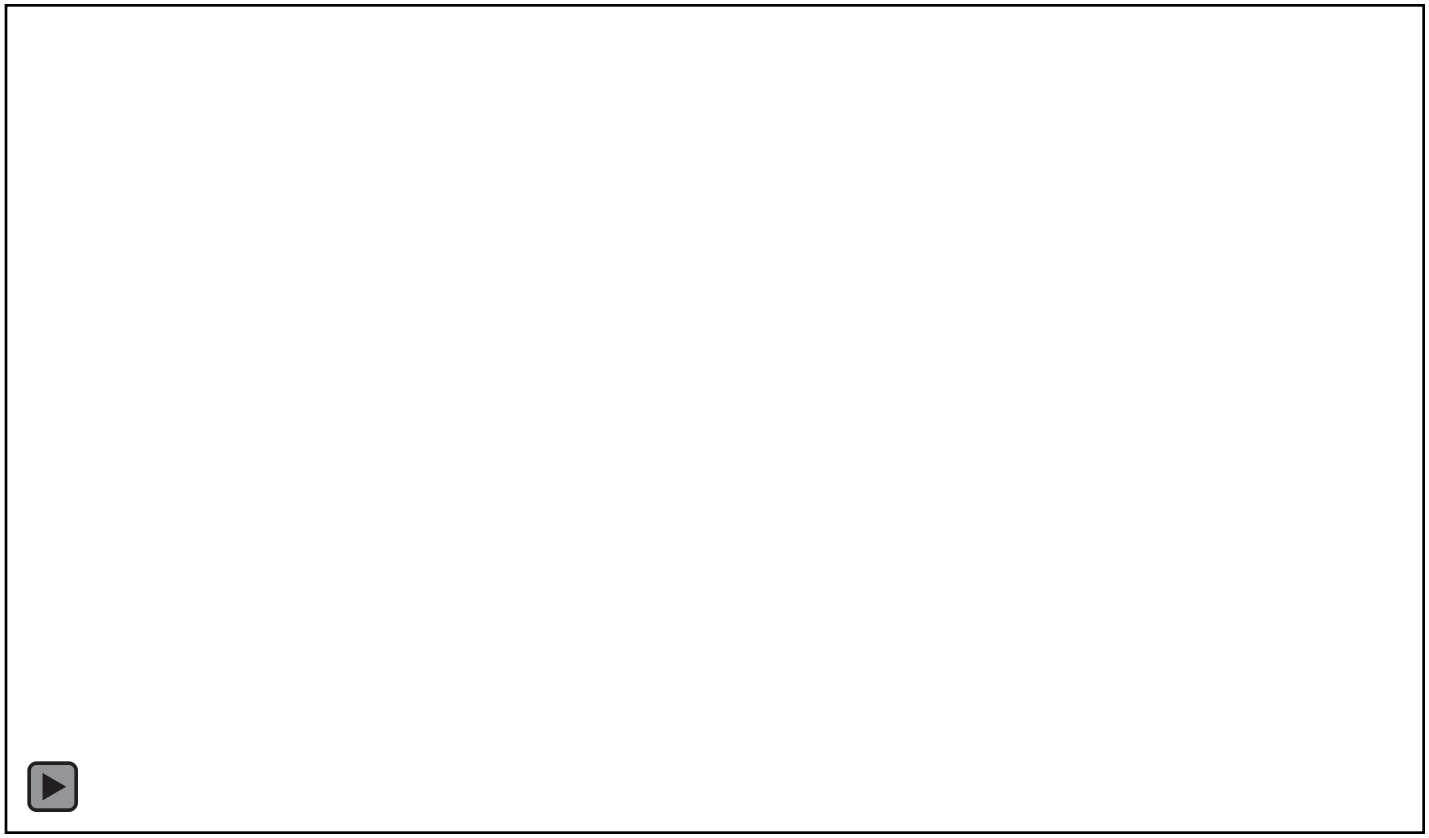
Observers and judges

- **Example: "Simple Persuasiveness Rating Scale"**

Persuasiveness is the capability of a person or argument to convince or persuade someone to accept a desired way of thinking. You will watch a brief movie review in which a reviewer will argue that a movie is either worth watching or is not worth watching. Use the following scale to indicate how persuasive you found the movie review to be overall.

1	2	3	4	5
Not at all persuasive	Slightly persuasive	Moderately persuasive	Very persuasive	Extremely persuasive

Observers and judges



https://youtu.be/RMnmddplxPo

Observers and judges

- How could we improve this judgement rating scale?
- What are other indicators of persuasiveness?
  - *How convincing did you find this movie review?*
  - *How much did you agree with the movie reviewer's arguments?*
  - *How much do you want to see the movie that was reviewed?*
- What are some other similar/dissimilar constructs?
  - *How likeable did you find the movie reviewer?*
  - *How similar to you was the movie reviewer?*
  - *How much did you enjoy the movie review?*

Observers and judges

- Understanding a measurement instrument*
  - Are measurements made using categories or dimensions?
    e.g., Was the video happy or sad? How positive was the video?
  - How much inference/abstraction/subjectivity is needed?
    e.g., Did the woman smile? Did the woman feel proud?
  - How (and how often) are measurements made?
    e.g., Events or intervals? Once per video? Once per minute?

*These and other aspects will be explored in your reading this week.

Observers and judges

- Assumptions of observational and judgement data
  - The stimulus is rich enough to inform measurement
  - Observers can generalize to new examples
- Pros and cons of observational/judgement data
  - Avoids many biases introduced by self-report methods
  - Uses similar information as affective computing algorithms
  - Can be very inefficient and subjective (when done poorly)
  - Can't access internal states of the stimulus participants

Observers and judges

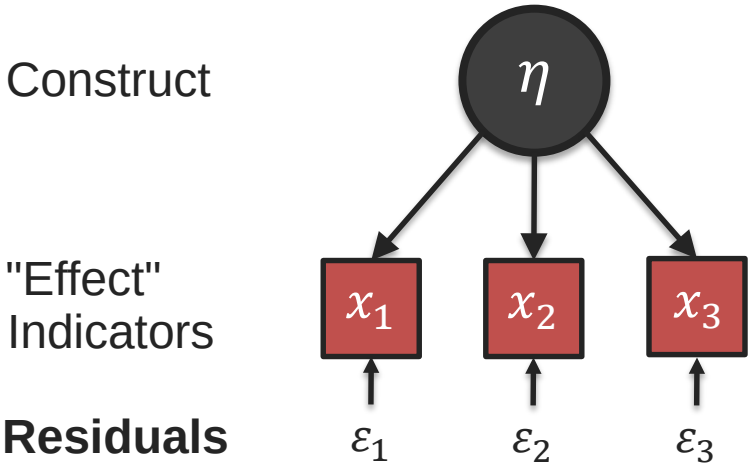
- Final notes on observers and judges
  - Provide some basic training (definitions and instructions)
  - Try to measure each construct in several different ways
  - Larger time-scales make measurement easier and faster
  - More difficult and subjective tasks require more observers
- Any questions about observers and judges?

When is measurement trustworthy?

Validity and reliability of measurement

Validity and reliability of measurement

- **Measures of indicators should ideally be influenced by the construct that causes them and nothing else**
- In this kind of latent variable model, the *residuals* contain all of the non-construct variance for each indicator (improving the scores)
- Sum and average aggregates assume there is *no* residual variance

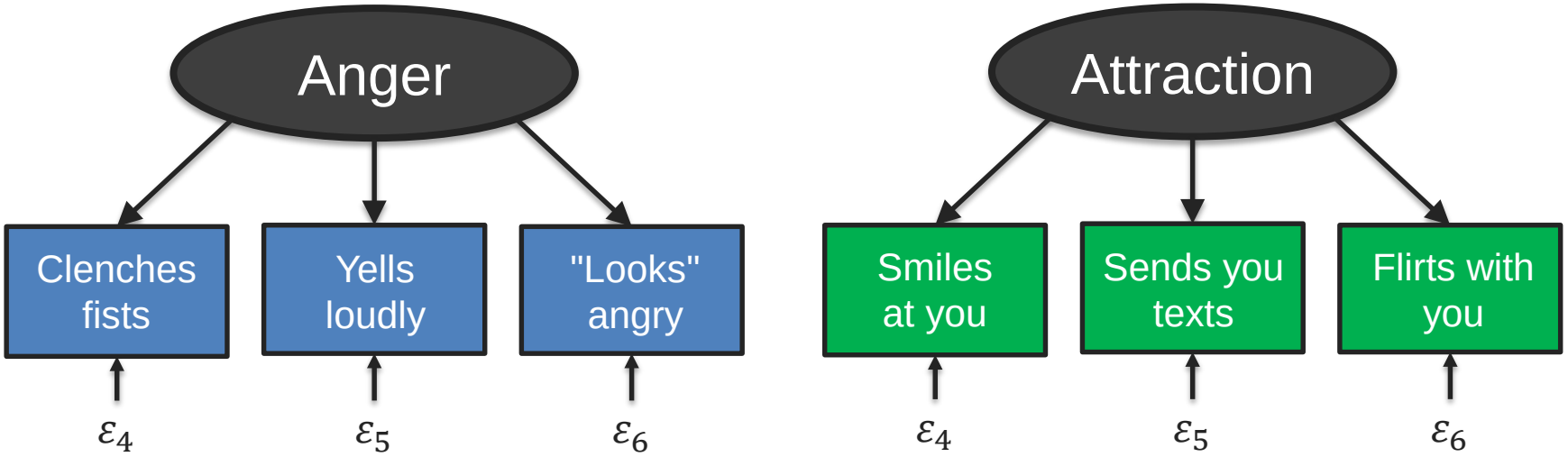


Validity and reliability of measurement

- **What else can influence a measurement?**
  - Other constructs that we did not intend to measure
  - Aspects of the measurement situation (who, when, where, how)
  - Stochastic processes that add random noise to measurements
- **What else can go wrong in measurement?**
  - The selected indicators can poorly represent the construct
  - The label or name assigned to the score can be misleading
  - Scores can be interpreted and/or ultimately used incorrectly

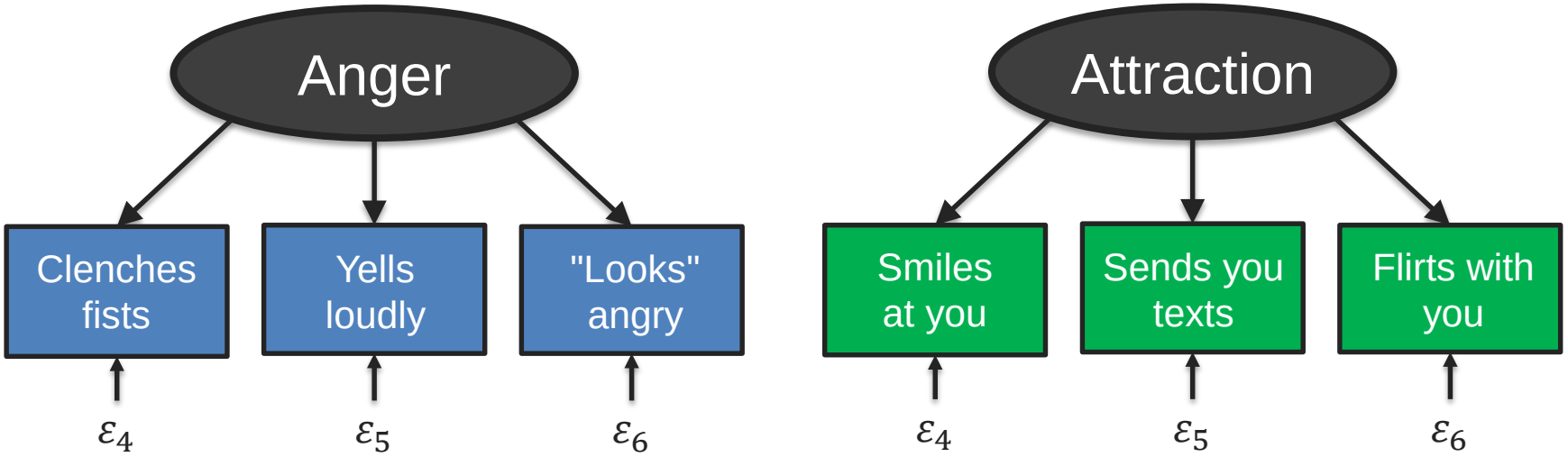
Validity and reliability of measurement

- What other constructs might influence these measures?
- What aspects of the measurement situation might matter?
- Could stochastic processes add noise to these measures?



Validity and reliability of measurement

- Do these indicators represent their constructs well?
- Are these good names/labels for the aggregate scores?
- How might these scores be interpreted or used incorrectly?



Validity and reliability of measurement

- **Validity** is a broad term for the overall connection between scores and the constructs they measure
- **Reliability** is specifically robustness to changes in the measurement situation (e.g., who, when, where, how)
- *Measurement is trustworthy to the extent that there is convincing evidence of its validity (and reliability)*
- **Validation** is the process of gathering and presenting evidence of measurements' validity (and reliability)

Validity and reliability of measurement

More on the reliability of measurement

- Do scores change along with the measurement context?
  - **Who:** different populations, observers, judges, interpreters
  - **When:** different occasions, times of day, temporal contexts
  - **Where:** different settings, locations, environmental conditions
  - **How:** different items, versions, response options, instructions
- Scores should be influenced by *constructs*, not context
- Scores with low reliability are thus noisy and/or biased
- It is hard to make good decisions using such scores

Validity and reliability of measurement

Examples of low reliability in psychology or medicine

- One doctor thinks a patient is depressed, but another doctor doesn't
- Participants report higher life satisfaction on Friday than on Monday
- Participants behave differently in the lab than they do at home
- A participant reports feeling "mad" (4/5) but not "angry" (0/5)

Examples of low reliability in affective computing

- A facial recognition algorithm performs worse on men with beards
- An emotion recognition algorithm performs worse during speech
- A "drowsy driver" algorithm performs worse outside of America

An introduction to validation

An introduction to validation

- Validation is a very complex and fascinating topic
- It is especially important in high-stakes areas
  - Medicine, criminal justice, finance, hiring, etc.
- There are three main phases of validation
  - **Substantive phase:** define the construct, select indicators
  - **Structural phase:** assess the indicator measures and reliability
  - **External phase:** assess how scores relate to other variables

An introduction to validation

- **Substantive phase of validation**
  - Do the selected indicators represent the construct well?
  - Would experts agree that these are a good set of indicators?
  - Are the methods used to measure the indicators reasonable?
  - *Literature review* and *construct conceptualization* are helpful
    - How will you define the construct? How have others?
    - Which indicators will you include? Are they representative?
    - How is the construct similar/dissimilar to other constructs?

An introduction to validation

- **Structural phase of validation**
  - Are the indicator measurements distributed as expected?
  - Does the pattern of indicator correlations match expectations?
  - Do the measurements show evidence of reliability?
  - *Latent variable models* and *reliability analyses* are helpful
    - Are all of the indicators inter-correlated? Are some outliers?
    - Do subsets of indicators cluster into facets (sub-constructs)?
    - Are scores consistent over time, across situations, etc.?

An introduction to validation

- **External phase of validation**
  - Do the scores correlate with other measures of the construct?
  - Are the scores unrelated to measures of different constructs?
  - Do the scores aid in the prediction of important outcomes?
  - Do the scores help differentiate between "known" groups?
  - *Latent variable models* and *linear statistical models* are helpful

Further Reading

Assigned:

- Girard, J. M., & Cohn, J. F. (2016). A primer on observational measurement. *Assessment*, *23*(4), 404–413. https://jmgirard.com/pubs/girard2016a/
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10/gbf8nx

Optional:

- Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, *15*(4), 380–387. https://doi.org/10/bnn2s3
- Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. https://doi.org/10/f9w6ff